클러스터링 알고리즘 기반의 임베딩 기법 성능 비교 및 분석

박정민, *박희민, **양선아, ***순위샹, ****이용주
경북대학교
pjm9562@naver.com, *gmlals9628@naver.com, **amy1353@naver.com,
***syx921120@gmain.com, ****yongju@knu.ac.kr

Performance Comparison and Analysis of Embedding methods based on Clustering Algorithms

Jungmin Park, *Heemin Park, **Seona Yang, ***Yuxiang Sun, ****Yongju Lee
Kyungpook National University

요 약

최근 구글, 아마존, LOD 등을 중심으로 지식 그래프(Knowledge graph)와 같은 검색 고도화 연구가 활발히 수행되고 있다.그러나 대규모 지식 그래프 인덱싱 시스템에서 데이터가 어떻게 임베딩(embedding)되고, 딥러닝(deep learning) 되는지는 상대적으로 거의 연구가 되지 않고 있다. 이에 본 논문에서는 임베딩 모델에 대한 성능평가를 통해 데이터셋에 대해 어떤 모델이 가장 좋은 지식 임베딩 방법을 도출하는지 분석한다.

1. 서론

최근 “한국판 뉴딜 정책”의 대표 과제인 “데이터 댐” 사업을 통해 공공 데이터 개방 및 AI 학습용 데이터 구축을 촉진하고 있다. 이에 구글, 아마존, LOD 등을 중심으로 지식 그래프(Knowledge graph)와 같은 검색 고도화 연구가 활발히 수행되고 있다. 그러나 현재까지 시맨틱 검색 시스템에 이러한 기술을 적용한 예는 거의 찾아보기 힘들다. 지식 임베딩(Embedding) 기법들을 적용한 연구는 현재 상당한 진전이 이루어졌다. 하지만 현재까지 시맨틱 검색 시스템에 이러한 기술을 적용한 예는 찾아보기 힘들다.

따라서 지식 임베딩을 기반으로 하는 시맨틱 검색 시스템을 구축함에 있어서 중요한 클러스터링 알고리즘 기반의 임베딩 기법에 대한 성능 비교를 본 논문에서 다룰 것이다. 임베딩 기법의 성능은 곧 시맨틱 검색 시스템의 성능과 직결되므로 각 기법들의 성능비교는 아주 필수적인 부분이다. 본 연구에서는 Translational distance model 의 TransD, TransR 과 Semantic matching model 의 RESCAL, DistMult. Deep learning model 의 ConvE 까지 총 5 개의 모델의 성능비교를 진행하였다.

본 연구를 통하여 최근 대형 IT 기업에서 활발히 연구중인 딥러닝 기반 시맨틱 검색 시스템의 발전[1]에 이바지할 수 있다. 더 나아가 기존 시스템보다 성능이 우수한 시맨틱 검색 시스템을 만들어 신기술 주도적인 역할을 수행할 수 있는 계기를 마련한다.

2. 관련 연구

2.1 지식 임베딩 방법의 분류

임베딩의 목적은 지식 그래프의 내부 구조를 유지 시키고 연속적인 벡터 공간에 주어, 목적어, 서술어 데이터들의 연산을 단순화하는 것이다. 여기서 지식그래프는 다양한 소스로부터 축적한 문장이나 단락에 기술된 주제를 파악하고 이를 대상으로 검색한 정보를 사용하여 검색결과를 향상시키는 것이다. 임베딩은 벡터화된 데이터들로 데이터간 관계를 추출하고 비슷한 개체들끼리 분류하여 지식 그래프를 구축하는데 널리 사용된다. 해당 논문의 주요 목적은 인덱스 구조를 기반으로 임베딩 방법이 검색 기능에 미치는 영향을 조사하는 것이다. 임베딩 방법에는 주로 translational distance model 을 대표하는 TransE[2], TransD[3], TransR[4]가 있고, Semantic Matching model 에는

RESCAL[5]와 DistMult[6]이 대표적으로 있으며, Deep Learning Model 을 대표하는 ConvE[7]와 ConvKB[8] 이렇게 세가지 모델로 나뉜다.

그래프 임베딩 방법에 있는 Word2Vec[9]에서 영감을 받아, 지식 그래프 임베딩에 translation invariance 를 도입한 translational distance model 기반의 TransE 임베딩 모델이 제안되었다. 여기서 Translation invariance 는 input 의 위치가 달라져도 output 이 동일한 값을 갖는 것을 말한다. 하지만 TransE 는 대규모 지식 그래프 임베딩에서 큰 발전을 이루었지만 1-N, N-1, N-N 같은 복잡한 관계를 처리하는 데 여전히 어려움을 가지고 있다. Semantic Matching model 기반 DistMult 모델은 복잡한 매트릭스를 사용하는 대신 head 와 tail 엔티티에서 주대각선 성분 외 모든 성분이 0 의 값을 가지는 대각 행렬만을 사용하여 대칭 관계 매개변수의 수를 줄일 수 있다. ComplEx 는 대칭 및 비대칭 관계를 처리하기위해 실수 및 가상 부분같은 복잡한 임베딩을 활용하여 DistMult 를 확장한다. 이는 계산 복잡성을 감소시킬 뿐만 아니라 엔티티 표현 능력을 향상시킨다. Deep Learning Model 기반 ConvE 는 CNN(Convolutional Neural Networks)을 도입하여 주어, 목적어, 서술어의 잠재적인 의미 정보를 포착한다. 이미지에서 feature extractor 을 하는 Convolutional layer 은 semantic matching model 에 비해 지식 그래프의 특징을 더 잘 추출할 수 있다.

2.2 시맨틱 클러스터링 알고리즘

클러스터링은 대표적인 비지도 학습으로, 데이터들의 특성을 고려해 데이터 클러스터(Cluster)를 정의하고 해당 데이터 집단을 대표할 수 있는 중심점을 찾는 것으로, 데이터 마이닝의 한 방법이다. 여기서 비지도 학습이란 정답이 없는 상태에서 학습시키는 것을 말한다. 데이터의 숨겨진 특징이나 구조를 찾기 위해 주로 사용한다. 아래에서 소개할 두개의 알고리즘 역시 비지도 학습이다.

밀도기반 알고리즘(DBSCAN)[10]은 비선형 클러스터의 효과적인 클러스터링을 위해 이웃한 개체와의 밀도를 계산하여 클러스터링을 진행한다. 쉽게 말하자면, 어느 점을 기준으로 반경 x 내에 점이 n 개 이상 있으면 하나의 군집으로 인식하는 방식이다. 클러스터내의 데이터 개수가 불필요하고 잡음에 대한 강인성이 높다는 것이 이 알고리즘의 장점이다. 공간 데이터베이스에서 취약했던 기존 클러스터링 알고리즘보다 성능이 좋고 본 연구에서 진행한 데이터셋의 형태에 적합한 알고리즘이기 때문에 이 알고리즘을 사용하게 되었다.

K-평균 알고리즘은 주어진 데이터를 k 개의 클러스터(Cluster)로 묶는 알고리즘이다. 이때 평균(Means)는 중심점과 데이터들의 평균거리를 뜻한다. 정리하자면, 평균을 이용해서 k 개의 클러스터를 만드는 알고리즘이라고 할 수 있다. 위의 밀도기반 알고리즘(DBSCAN)과 달리 중심기반 알고리즘이다. k 값을 정하기 어렵고 비수치 데이터이거나 임의의 거리 측도로 계산해야 할 때는 사용할 수 없지만, 비교적 구현이 쉽고 구형(spherical)의 데이터에 적합하기 때문에 본 연구에서 K-평균 알고리즘을 사용하게 되었다.

3. 실험

3.1 실험 환경 및 평가 기준

객관적이고 현실적인 결과를 얻기위해 본 연구에서는 세계적으로 널리 사용되고 있는 리하이 대학교의 벤치마크 데이터셋인 Lehigh university benchmark(LUBM)을 사용했다. LUBM 데이터셋은 230,061 의 triple 과 38,334 개의 주어(subject), 17 개의 서술어(predicate), 29,635 개의 목적어(object)를 포함하고 있고, 총 36.7MB 사이즈이다.

평가기준으로는 종합적인 성능 평가 방법으로 가장 널리 사용되고 있는 재현율 R(Recall), 정확률 P(Precision), F-척도 F(F-measure)를 사용했다. 재현율은 실제 true 값인 것들 중 true 라고 예측된 값의 비율을 나타내는 것으로, 완전성을 측정하는 척도이다. 재현율을 통해 결과를 얼마나 잘 예측하는지 알 수 있다. 반면, 정확률은 true 값이라고 예측된 값 중 실제 true 값의 비율을 나타내는 척도로 정확도를 측정해준다. 정확률을 통해 얼마나 정확하게 예측했는지 알 수 있다. 대개 R 과 P 의 측정은 별개로 논의되지 않으며, R 과 P 의 값은 두 값의 조화 평균인 F 와 같은 단일 측정을 통해 결합시킬 수 있다. 조화평균을 통해 평균적인 변화율을 측정할 수 있다.

3.2 실험

우리는 translational distance model 모델인 TransnD[3] 모델과 TransR[4] 모델, Semantic Matching model 모델인 RESCAL[5] 모델과 DistMult[6] 모델, neural 딥러닝 모델인 ConvE[7] 모델에 대해서 시맨틱 클러스터링 성능 평가 실험을 진행하였다.

각 모델을 이용하여 LUBM 데이터셋을 벡터화 시킨 후, 밀도기반 클러스터링인 DBSCAN 과 구형으로 분포된 데이터에 적용가능한 K-means 알고리즘을 사용하여 군집화 하였다. 실험 결과는 표 1, 표 2, 그래프 1, 그래프 2 를 통해 나타내었다. 실험 결과, TransR 모델이 P, R, F 측면에서 가장 훌륭한 결과가 나왔다.

TransR 모델은 TransH 모델의 공간 투영 아이디어를 바탕으로 한 모델이며, 엔티티(head, tail)와 관계(relation)를 구분된 공간에 임베딩한다는 특징이 있다. 즉, 엔티티 공간에서 해당 관계 공간까지 엔티티를 투영(projection)하고 투영된

엔티티 간의 변환을 구축하는 방식으로 온톨로지 이질성 문제를 해결하는 모델이다. 이러한 TrnasR[8]의 특성 때문에 좋은 결과가 나온 것으로 보인다.

반면 TransD, RESCAL, DistMult 및 ConvE 는 모두 저조한 성과를 보인다. TransD 는 TransR 과 마찬가지로 엔티티와 관계를 서로 다른 공간에 임베딩 하지만, TransD 는 TransR 처럼 모든 임베딩에 대해 동일한 관계별 투영을 하는 것이 아니라 엔티티 관계별 투영 행렬을 사용한다. 그래도 TransR 과 마찬가지로 엔티티와 관계를 서로 다른 공간에 임베딩 하기 때문에 두번째로 좋은 결과가 나온 것을 확인할 수 있다. RESCAL[9] 모델은 관계형 임베딩 모델으로, 여러 행렬을 사용하여 엔티티 간의 관계를 나타낸다. RESCAL 은 복잡한 관계형 패턴을 포착할 수 있는 강력한 모델이지만, 임베딩 차원과 관련하여 2 차 런타임 및 메모리 복잡성이 있기 때문에 매우 큰 지식 그래프로 확장하기 어렵다. 이런 문제점 때문에 RESCAL 의 성과가 저조하다고 볼 수 있다.

DistMult 또한 매개변수 수를 제외하고는 RESCAL 과 유사한 모델이다. DistMult 는 RESCAL 과는 다르게 대각 행렬을 사용하여 관계 매개변수를 줄이기 때문에 RESCAL 보다 매개변수 수가 적게는 1/10 배에서 1/100 배로 줄어드는 효과를 볼 수 있지만, 대각 행렬을 사용한 단순한 관계식이기 때문에 대칭 관계만 모델링 할 수 있어 일반 지식 그래프에는 적합하지 않은 특징을 가지고 있다. 실험 결과에서도 DistMult 의 결과가 가장 좋지 않은 것으로 확인됐다. ConvE 는 합성곱(convolution) 레이어를 통과하고, 선형 변형을 통해 벡터로 변환하는 모델이다. 그러나 특성 벡터들 간의 상호작용 중 일부는 알아낼 수 없다는 특징을 가지고 있다. 해당 실험에서는 TransD, TransR 모델보다는 저조한 성능을 보이지만 RESCAL 모델과 DistMult 모델보다는 우수한 성능을 보이는 것을 확인할 수 있다.

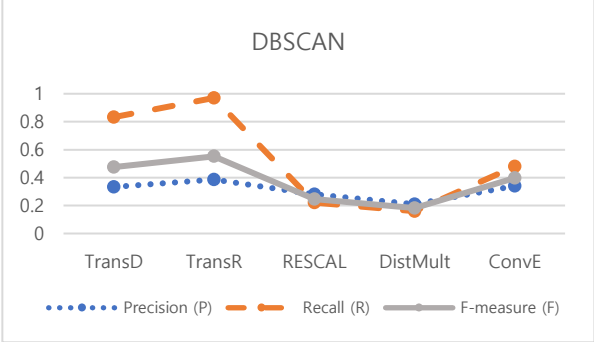
각각의 임베딩 모델들은 서로 다른 특성과 강점을 가지고 있고, 항상 특정 모델이 우수하다고 평가할 수는 없다. 그러나 LUBM 데이터셋에서는 엔티티와 관계를 구분된 공간에 투영하는 모델이 우수성을 보이고 그중 TransR 모델을 활용한 방법이 가장 성능이 좋다고 볼 수 있다.

4. 결론

최근 공공데이터 개방 및 AI 학습용 대용량 규모의 빅데이터 클라우드를 구축으로 지식그래프와 같은 검색 고도화 연구가 활발히 수행되고 있다. 하지만 현재까지 시맨틱 웹은 이질성 문제, 단어 간 상호작용 등 많은 문제에 직면해 있었다. 이에 유사한 단어일 수록 가까운 거리에 위치하도록 벡터 값을 가지는 단어 임베딩도 연구가 되었지만 지식 그래프는 트리플 구조로 저장되어야 하기 때문에 그래프 임베딩 기법이 필요했다. 따라서 본 논문에서 시맨틱 검색 시스템 환경에서의 지식 임베딩 기법에 대한 성능 비교 실험을 진행하였다.

Models	Precision (P)	Recall (R)	F-measure (F)
TransD	0.333	0.832	0.476
TransR	0.386	0.97	0.552
RESCAL	0.28	0.22	0.246
DistMult	0.21	0.16	0.182
ConvE	0.34	0.48	0.398

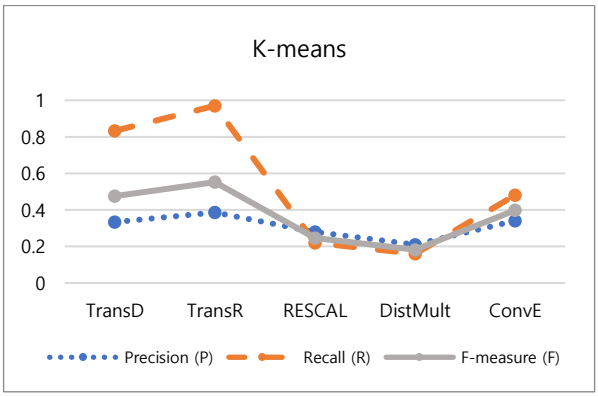
<표 1> DBSCAN 기반의 임베딩 성능



[그림 1] DBSCAN 기반의 임베딩 성능

Models	Precision (P)	Recall (R)	F-measure (F)
TransD	0.26	0.762	0.387
TransR	0.345	0.93	0.503
RESCAL	0.25	0.182	0.21
DistMult	0.334	0.287	0.308
ConvE	0.335	0.425	0.375

<표 2> K-means 기반의 임베딩 성능



[그림 2] K-means 기반의 임베딩 성능

재현율, 정확률, F-척도에 대해 TransD, TransR, RESCAL, DistMult, ConvE 의 총 5 가지 임베딩 모델에 대해 LUBM 데이터셋으로 실험을 진행한 결과 TransR 모델에서 가장 훌륭한 실험결과가 나왔다. 각각의 임베딩 모델들은 서로 다른 특성과 강점을 가지고 있고, 항상 특정 모델이 우수하다고 평가할 수는 없다. 그러나 LUBM 데이터셋에서는 엔티티와 관계를 구분된

공간에 투영하는 모델인 TransD 와 TransR 이 우수성을 보이고 그중 모든 임베딩에 대해 동일한 관계별 투영을 하는 TransR 모델을 활용한 방법이 가장 성능이 좋다고 볼 수 있다.

가장 최적의 임베딩 기법인 TransR 모델을 사용하여 좀 더 나은 성능의 시맨틱 검색 시스템을 구축할 수 있다. 그러나 지식 임베딩 기법의 지속적인 최적화가 요구되므로 최적화는 연구가 끝난 후에도 꾸준히 진행할 예정이다. TransR 로 임베딩 된 지식그래프에 좀 더 빠르고 정확하게 검색하기 위해서 인덱스 방법을 사용해야 한다. 따라서 인덱스의 구조에 대해서도 추가적인 연구가 필요하다. 향후 본 연구에서 개발된 시맨틱 시스템을 활용하여 '질의/응답 시스템'이나 '추천시스템' 개발로 그 활용분야를 넓힐 수 있다.

감사의 글

이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2016R1D1A1B02008553). 본 연구는 2021년 과학기술정통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음(2021-0-01082).

참고문헌

[1] 윤원준(Won Joon Yun), 정소이(Soyi Jung), 박지홍(Jihong Park), and 김중헌(Joongheon Kim). "딥러닝 기반 시맨틱 통신 기술 동향." 한국통신학회 학술대회논문집 2021.6 (2021): 228-229.

[2] Antoine Boreds, Nicolas Usunier, Alberto Garcia-Buran, Translating Embeddings for Modeling Multi-relational Data, Proc. CNRS. (2013).

[3] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu and Jun Zhao, Knowledge Graph Embedding via Dynamic Mapping Matrix, Proc. 53rd Annual Meeting of the Association for Computational Linguistics. (2015), 687-696.

[4] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu and Xuan Zhu, Learning Entity and Relation Embeddings for Knowledge Graph Completion, Proc. the National Conference on Artificial Intelligence, (2015). 2181-2187.

[5] Maximilian Nickel, Volker Tresp and Hans-Peter Kriegel, A Three-Way Model for Collective Learning on Multi-Relational Data Maximilian, Proc. 28th International Conference on Machine Learning. (2011), 809-816.

[6] Bishan Yang1, Wen-tau Yih, Xiaodong He, Jianfeng Gao and Li Deng, Embedding entities and relations for learning and inference in knowledge bases, Proc. the International Conference on Learning Representations. (2015), 1-12.

[7] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp and Sebastian Riedel, Convolutional 2D Knowledge Graph Embeddings, Proc. 32nd AAAI Conference on Artificial Intelligence. (2018), 1811-1818.

[8] Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen and Dinh Phung, A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network, Proc. the North American Chapter of the Association for Computational Linguistics. (2018), 327-333.

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, Efficient estimation of word representations in vector space, Proc. ICLR Workshop. (2013), 1-13.

[10] Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.(1996),1 - 6